Rejection Criteria Based on Outliers in the KiDS Photometric Redshifts and PDF Distributions Derived by Machine Learning

Valeria Amaro, Stefano Cavuoti, Massimo Brescia, Giuseppe Riccio, Crescenzo Tortora, Maurizio D'Addona, Michele Delli Veneri, Nicola R. Napolitano, Mario Radovich, and Giuseppe Longo

Abstract The Probability Density Function (PDF) provides an estimate of the photometric redshift (zphot) prediction error. It is crucial for current and future sky surveys, characterized by strict requirements on the zphot precision, reliability and completeness. The present work stands on the assumption that properly defined rejection criteria, capable of identifying and rejecting potential outliers, can increase the precision of zphot estimates and of their cumulative PDF, without sacrificing much in terms of completeness of the sample. We provide a way to assess rejection through proper cuts on the shape descriptors of a PDF, such as the width and the height of the maximum PDF's peak. In this work we tested these rejection criteria to galaxies with photometry extracted from the Kilo Degree Survey (KiDS) ESO Data Release 4, proving that such approach could lead to significant improvements to the zphot quality: e.g., for the clipped sample showing the best trade-off between precision and completeness, we achieve a reduction in outliers fraction of $\simeq 75\%$ and an improvement of $\simeq 6\%$ for NMAD, with respect to the original data set, preserving the $\simeq 93\%$ of its content.

V. Amaro (✉) · N. R. Napolitano
School of Physics and Astronomy, Sun Yat-sen University, Zhuhai Campus, Guangzhou 519082, People's Republic of China
e-mail: valeriaa@mail.sysu.edu.cn

S. Cavuoti · M. Brescia · G. Riccio · C. Tortora
INAF - Astronomical Observatory of Capodimonte, Salita Moiariello 16, 80131 Napoli, Italy
e-mail: stefano.cavuoti@gmail.com

M. D'Addona · G. Longo
Department of Physics, University of Naples Federico II, Strada Vicinale Cupa Cintia, 21, 80126 Napoli, Italy
e-mail: longo@na.infn.it

M. D. Veneri
DIETI, University of Naples Federico II, via Claudio 21, 80125 Napoli, Italy
e-mail: micheledelliveneri@gmail.com

M. Radovich
INAF – Osservatorio Astronomico di Padova, Vicolo Osservatorio 5, 35122 Padova, Italy
e-mail: mario.radovich@inaf.it

1 Introduction

Photometric redshifts (zphot) are crucial for modern cosmology surveys, since they provide the only viable approach to determine the distances of large samples of galaxies. Over the years they have been used to constrain the dark matter and dark energy contents of the Universe through weak gravitational lensing [1–3], to reconstruct the cosmic Large Scale Structure [4], to identify galaxy clusters and groups [5–7], to disentangle the nature of astronomical sources [8, 9]; to map the galaxy colour-redshift relationships [10] and to measure the baryonic acoustic oscillations spectrum [11, 12].

Baum [13] first noticed that the stretching of a galaxy spectrum due to the redshift affects the observed colours and hence, if the correlation between photometry and redshift can be uncovered, multi-band photometry could become a powerful tool to estimate redshifts.

However it became immediately apparent that such correlation is highly non-linear and too complex to be derived analytically [14] and that the derivation of zphot required alternative, interpolative approaches.

Nowadays, it is common praxis to divide these methods into two broad classes: the Spectral Energy Distribution (SED) template fitting methods (e.g., [15–17]) and the empirical (or interpolative) methods (e.g., [18–28]), both characterized by their pros and cons.

SED methods rely on fitting the multi-wavelength photometric observations of the objects to a library of synthetic or observed template SEDs, shifted to create synthetic magnitudes for each galaxy template as a function of the redshift. SED fitting methods, while relying on many assumptions, allow pushing zphot beyond the spectroscopic limit.

Empirical methods use instead an a priori spectroscopic knowledge (zspec) for a subsample of objects to infer the complicated relationship existing between the photometric data (i.e. magnitudes and or derived colours, in some cases complemented by morphological information) and the redshift. Among these methods, a critical role is played by machine learning (ML). Among the many ML models applied to the zphot estimation, we quote just a few: neural networks, boosted decision trees, random forests, self-organized maps, convolutional neural networks (see [29] and references therein). A primary advantage of ML is the high accuracy of predicted zphot within limits imposed by the spectroscopic knowledge base (KB). On the other hand, ML methods have an inferior capability to extrapolate information outside the regions of the parameter space properly sampled by the training data and, for instance, they cannot be used to estimate the redshift of objects fainter than those present in the spectroscopic sample.

Extensive reviews of both approaches can be found in [30–32]. An additional difference between the two approaches is that SED fitting methods allow obtaining, at once, the zphot, the spectral type of the objects and the Probability Density Function (hereafter, PDF) of the predicted zphot. In contrast, ML-based methods do not naturally provide a PDF unless special procedures are implemented.

The essential complementarity of the two methodologies was proven to be the most reliable and efficient way to produce a high-quality zphot catalogue [33], particularly

suitable for extensive surveys, like Euclid (Desprez et al., in prep.) and VRST [34]. In this work, we explore the possibility to improve the quality of zphot predictions by excluding from the data potential outliers without loosing much in completeness.

The work is structured as follows: in Sect. 2, we introduce some aspects of PDF evaluation; in Sect. 3, we describe the photometry and spectroscopy used for the analysis herein; in Sect. 4, we give a description of the method for the rejection criteria identification as well as of the statistical estimators used to quantify the performance of both zphot and cumulative PDF statistics. In Sect. 5, we show all the results, and, finally, in Sect. 6, we draw the conclusions.

2 Probability Density Function

In general terms, a PDF is a way to parametrise the uncertainty on the zphot prediction and to provide a robust estimate of the reliability of any individual redshift. From a rigorous statistical point of view, however, a PDF is an intrinsic property of a particular phenomenon, regardless of the measurement methods that allow quantifying the phenomenon itself [35].

Unfortunately, in the zphot context, the PDF depends both on the measurement methods (and chosen internal parameters of the methods themselves) as well as on the underlying physical assumptions. In this sense, the definition of a PDF in the context of zphot estimation needs to be taken with some caution [36].

The factors affecting the reliability of zphot PDFs are: photometric errors, intrinsic errors of the methods and statistical biases. The PDF becomes, therefore, just a way to somehow compress the information contained in a single error estimate. In other words, the parametrization of a single error, through a probability, allows to cover an entire redshift range (with the chosen bin accuracy), thus leading to an increase of the information rate in order to match the precision required by a specific scientific goal (cf. for instance, the cases of the determination of cosmological parameters [37], and weak lensing measurements [38]).

Therefore, over the last few years, much attention has been paid to develop methods able to compute a full zphot PDF for both individual sources and entire galaxy samples [39–45].

The study of the PDFs and their properties (see Sect. 4) represents a useful tool to test zphot reliability. In fact, rejection criteria, aimed at removing unreliable zphot and PDFs estimates, as long as they are reproducible in the photometric space, can improve the precision of the results in terms of both NMAD and fraction of outliers for zphot estimates, as well as the quality of individual PDFs and their cumulative performances described by the statistical indicators discussed in Sect. 4. Of course, this comes at the cost of a decreased completeness.

3 Data

As spectroscopic knowledge base we used zspec for $136, 057$ galaxies extracted from the fourth Data Release (DR) of the ESO Public Kilo-Degree Survey (hereafter, KiDS-ESO-DR4, [46]) which combines data from KiDS and the VISTA Kilo degree INfrared Galaxy survey (VIKING; [47]).

KiDS is an optical survey of about 1350 deg^2 carried in 4 bands (ugri) with limiting magnitude r=25.0 AB (5 σ in 2"), i.e. 2.5 magnitudes deeper than the Sloan Digital Sky Survey (SDSS), in good seeing conditions (~0.7" median full width at half-maximum, FWHM, in the r band). KiDS has been complemented with the NIR photometry in the five bands Z, Y, J, H and Ks from the VIKING survey.

The survey is complemented by a set of spectroscopic observations available within the KiDS collaboration as well as from other surveys: COSMOS [48], zCOSMOS [49], CDFS [50], DEEP2 [51] and GAMA DR2 and DR3 [52, 53] fields public data.

The photometry used in this work consists of the 9 Gaussian Aperture and PSF (GAaP) magnitudes (u, g, r, i, Z, Y, J, H, Ks), corrected for extinction and zero-point offsets, and 8 derived colours, for a total of 17 photometric parameters for each object. After investigating the photometric parameter distribution for the sample, the data was cleaned by cutting the tails of the magnitude distributions in order to ensure a homogeneous distribution of training points in the parameter space.

The spectroscopic data set has been randomly shuffled and split in a 70% training set and a 30% test set (95, 261 and 40, 796 sources, respectively). We stress that objects in the test set used to evaluate and validate the trained model are not used during the training phase of the methods (see Sect. 4). In what follows, we used PDFs obtained using METAPHOR (Machine-learning Estimation Tool for Accurate PHOtometric Redshifts [54]) on KiDS-ESO-DR4 data (cf. [36] for details).

METAPHOR is a modular workflow designed to produce both zphot and related PDFs. The internal zphot estimation engine is MLPQNA (Multi-Layer Perceptron trained with Quasi-Newton Algorithm; [55, 56]) while for the zphot we used the *best-estimates* as defined in [36].

We just recall that the selected binning step in zphot used to estimate individual PDFs is 0.02 and the spectroscopic depth available is equal to zspec=7.01. The zspec distribution for the whole test set is shown in Fig. 1.

4 Methods

The zphot statistics are calculated on the residuals:

$$\Delta z = (z_{spec} - z_{phot})/(1 + z_{spec}) \tag{1}$$

using as zphot the *best-estimate* referenced in Sect. 3.

We use as accuracy estimators the mean (or bias), the fraction of catastrophic outliers, defined as those objects for which $|\Delta z| > 0.15$, and the normalized median absolute deviation (NMAD), defined as:

$$NMAD = 1.4826 \times median(|\Delta z - median(\Delta z)|) \tag{2}$$

The shapes of the PDFs, and therefore their intrinsic quality, can be characterized in terms of:

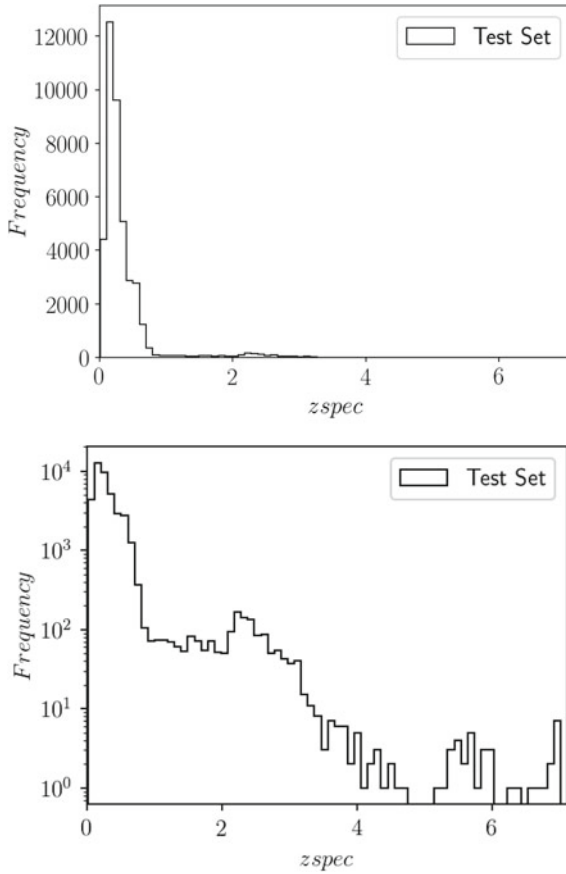- *pdfWidth*: the width of the PDF in terms of redshift;

Fig. 1 The zspec distribution of the initial test set (40, 796 sources). The bottom diagram is the same top plot but in a logarithmic scale

- *pdfNBins*: the total number of bins of chosen amplitude (which defines the accuracy of the PDF itself), in which the PDF is different from 0;
- *pdfPeakHeight*: the amplitude of the peak of the PDF, i.e. the value of the maximum probability of the PDF.

The cumulative performance of the stacked PDF on the entire sample is instead evaluated by means of the following three estimators:

- $f_{0.05}$: the percentage of residuals Δz within ± 0.05;
- $f_{0.15}$: the percentage of residuals Δz within ± 0.15;
- $\langle \Delta z \rangle$: the average of all the residuals Δz of the stacked PDFs.

where by stacked PDFs we mean the individual zphot PDFs transformed into the PDFs of scaled residuals Δz and then stacked for the entire sample.

Furthermore, the quality of the individual PDFs is evaluated against the single corresponding zspec in the test set, by defining five types of occurrences:

- *zspecClass* = 0: the zspec is within the *bin* containing the peak of the PDF;
- *zspecClass* = 1: the zspec falls in one bin from the peak of the PDF;
- *zspecClass* = 2: the zspec falls into the PDF, e.g. in a bin in which the PDF is different from zero;
- *zspecClass* = 3: the zspec falls in the first bin outside the limits of the PDF;
- *zspecClass* = 4: the zspec falls out of the first bin outside the limits of the PDF.

Finally, we use two additional diagnostics to analyze the *cumulative* performance of the PDFs: the credibility analysis presented in [57] and the Probability Integral Transform (hereafter PIT), described in [58].

The credibility test should assess if PDFs have the correct *width* or, in other words, it is a test of the *overconfidence* of any method used to calculate the PDFs. In particular, the method is considered overconfident if the produced PDFs result too narrow, i.e. too sharply peaked, underconfident otherwise. The implementation of the credibility method is straightforward and is reached by computing the threshold credibility C_i for the i-th galaxy with

$$C_i = \sum_{z \in p_i \geq p_i(z_{spec,i})} p_i(z) \tag{3}$$

where p_i is the normalized PDF for the i-th galaxy. The credibility is then tested by calculating the cumulative distribution $F(C)$, which should be equal to C. $F(C)$ is a q-q plot, (a typical quantile-quantile plot used to compare two distributions), in which F is expected to match C, i.e. it follows the bisector in the F and C ranges equal to [0,1]. Therefore, the *overconfidence* corresponds to $F(c)$ falling below the bisector (implying that too few galaxies have zspec with a given credibility interval), otherwise, the *underconfidence* occurs. In both cases, this method indicates the inaccuracy of the error budget [57]. *Overconfidence* and *underconfidence* are plotted in Fig. 2.

The PIT analysis measures how consistent are the predicted zphot and the true redshift (zspec) distributions, by calculating the histogram for the following probabilities:

$$p_i = F_i(x_i) \tag{4}$$

F_i in Eq. 4 is the cumulative distribution function (CDF) of the i-th object and $x_i = zspec_i$. The closer the histogram is to a uniform distribution, the better is the calibration between zphot and zspec distributions. A strongly U-shaped PIT histogram denotes a highly *underdispersive* character of the zphot distribution.

The visual inspection of a PIT can, therefore, shed light on the consistency between the zspec and zphot distributions. In particular, if the PDFs are too broad, then the relative PIT histogram appears overdispersed, that is with a peak in the centre of the histogram itself. In contrast, if the PDFs are too narrow, then the PIT is U-shaped and it results underdispersed. Finally, only when the widths of the PDFs agree with the
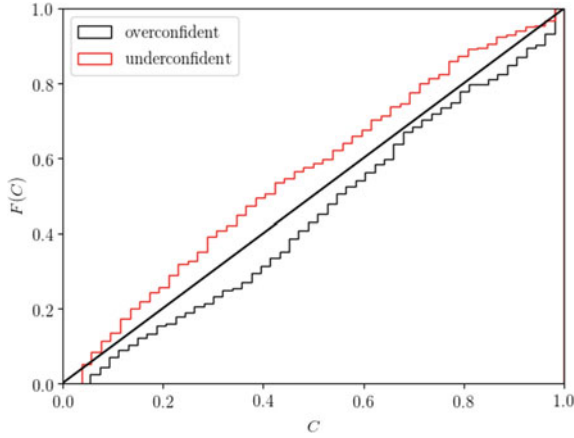
Fig. 2 Credibility analysis examples of *overconfidence*, represented by the black curve below the bisector of the plot $F(C)$ *vs* C, and of *underconfidence*, indicated by the red curve above the same bisector
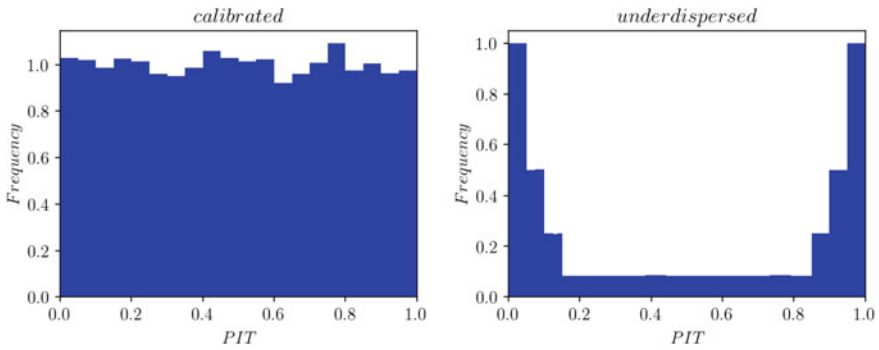

Fig. 3 Examples of well-calibrated (left) and underdispersed U-shaped PIT (right). The overdispersion is simply the opposite of the right panel, with a very broad PDF distribution, having a high peak in the centre of the diagram

discrepancies between zphot and zspec, then a uniformly distributed PIT histogram is produced. In Fig. 3 an example of well-calibrated and *underdispersed* PIT histograms is shown.

Credibility and PIT tests are complementary since both the underdispersion and the overconfidence are related to the narrowness of the PDFs. The narrower the PDFs are, the more the PIT histogram is underdispersed and the results of credibility are overconfident.

5   Results

The initial test dataset was composed by 40, 796 zphot estimates and relative individual PDFs. Among these, we have 970 outlier sources (2.4%), and 39, 826 non-outliers

(97.6%): outliers were singled out as explained in Sect. 4. We then proceeded with
the visual inspection of the individual PDFs, i.e. their width, number of bins, the
height of the maximum peak. To do so, we first derived a set of statistical descriptors, such as mean, standard deviation and the minimum and maximum values of
the PDF shape properties, dividing the sample into outliers and non-outliers. These
values are given in Tables 1 and 2, respectively. By comparing the mean values of
such descriptors, the expected differences for the two populations become apparent: outliers have wider PDFs, with a higher number of bins (intervals of amplitude
$\Delta z = 0.02$), in which the PDF is not null and lower peaks with respect to those for
non-outliers samples.

Table 1 Statistics of the three descriptors of the PDF shape: width, number of bins, and maximum peak height, defined in Sect. 4, for the outliers in the test set

PDF feature	Mean	σ	Min	Max
PdfWidth	3.94	2.33	0.1	7.02
PdfNBins	124.90	71.74	5	277
PdfPeakHeight	0.067	0.072	0.011	0.66

Table 2 Statistics of the three descriptors of the PDF shape: width, number of bins, and maximum peak height, defined in Sect. 4, for the non-outliers in the test set

PDF feature	Mean	σ	Min	Max
PdfWidth	1.23	1.82	0.040	7.02
PdfNBins	25.01	29.91	2	305
PdfPeakHeight	0.28	0.15	0.012	0.99

For the test set data, in Figs. 4, 5 and 6 we plot, respectively: *(i)* the histogram of the
pdfWidth distribution, *(ii)* the scatter plot of *PdfPeakHeight* against *PdfWidth*, and
(iii) PdfNBins versus *PdfWidth*, distinguishing outliers and non-outliers populations.
    The inspection of these plots led us to define four data sets:

- In Fig. 4, we can see that a cut of objects with *PdfWidth* > 4 can remove a fraction
  of outliers $\simeq 1.1\%$ from the test sample. We, therefore, define a first (*Cut-1*) data
  set of objects with reliable PDF widths, using the condition:

$$pdf Width < 4 \qquad (5)$$

with which we come out with 36, 170 sources (of which 518 are outliers, and
35, 652 non-outliers, respectively, 1.4% and 98.6% of the total number of objects
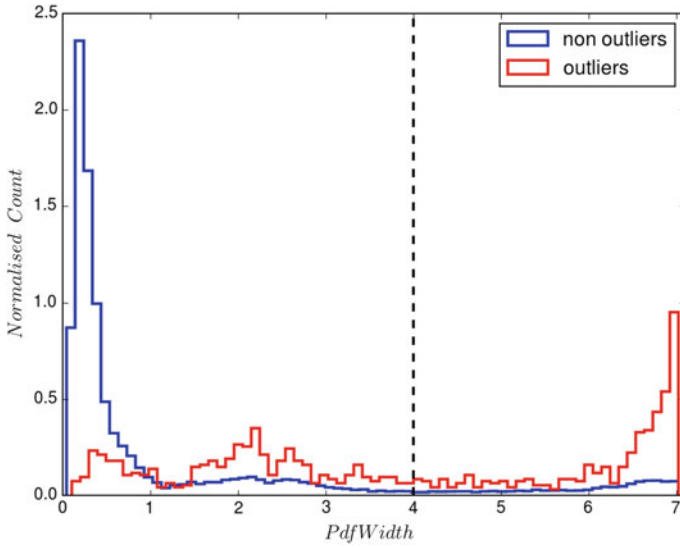in the sample).

Fig. 4 *PdfWidth* normalized counts for outliers and non-outliers populations. Dashed vertical line: *PdfWidth* value equal to 4, identified as threshold for clipping outliers that populate the region with widths larger than 4 (*Cut-1*)
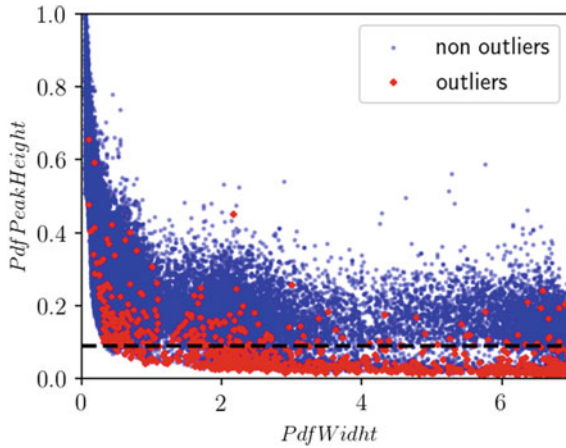

Fig. 5 Scatter plot of *PdfPeakHeight* versus *PdfWidth*. The dashed horizontal line indicates the *PdfPeakHeight* value, equal to 0.09, identified as threshold for removing outliers laying under the line (see *Cut-2* in the text)

- In Fig. 5, we show the *PdfPeakHeight* versus the *PdfWidth*. Outliers lay at the bottom of the scatter plot, thus allowing to define a second data set (*Cut-2*) via the condition:
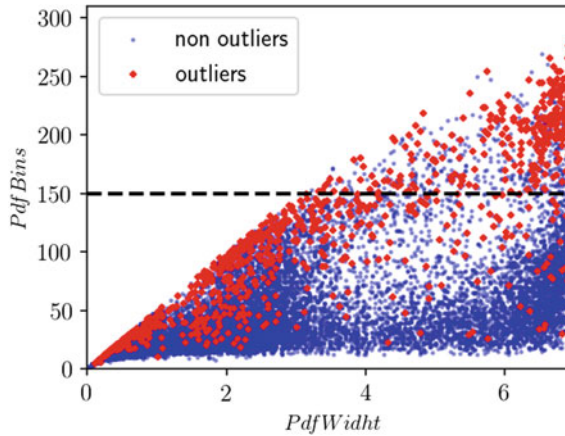
$$PdfPeakHeight > 0.09 \tag{6}$$

Fig. 6 Scatter plot *PdfNBins* versus *PdfWidth*. The dashed horizontal line identifies the value of *PdfBins* equal to 150, useful to clip the outliers populating the region above this threshold (see *Cut-3* in the text)

The *Cut-2* data set contains a total of 38, 107 sources: 226 outliers (0.6%), and 37, 881 non-outliers (99.4%).

- In Fig. 6, we show the distribution of the descriptors *PdfBins* vs *PdfWidth*. As expected, the majority of the outliers rests in the region with higher values of PDF width and number of bins. This allows defining a third data set *Cut-3*, by rejecting objects with *PdfBins* > 150. This third data set consists of 39, 905 sources, of which 591 are outliers (1, 5%) and 39, 314 non-outliers (98.5%).

- Finally, we derived a fourth data set (*Cut-4*) via the combination of *Cut-1* and *Cut-2*. This last data set contains 34, 802 sources, of which 196 (0.6%) are outliers and 34, 606 (99.4%) non-outliers.

Some additional tests showed that making more severe cuts would result in an uncomfortable loss in completeness. For instance, by producing an additional data set *Cut-5*, selecting objects with *PdfWidth* < 1, having less than 150 bins and a maximum peak height of at least 0.15, we reduced the test set by 34% (27, 297 out of 40, 796 sources).

5.1 Zphot and Stacked PDF Statistics

The results in terms of both zphot statistics and of *cumulative* PDF performance are reported in Table 3, for the test set and the four data sets corresponding to the different cuts. As we can see, the NMAD statistics is not different for the four adopted cuts, and there is only a slight improvement with respect to the whole data set. It has to be noted, however, that all cuts prove quite effective in reducing the fraction of outliers,

with an improvement of 39.8%, 75.0%, 37.7%, and 76.3% (for *Cut-1*, *Cut-2*, *Cut-3*, and *Cut-4*, respectively) with respect to the test set.

In the case of the data set extracted from *Cut-5*, we were left with a fraction of outliers $\simeq 0.2\%$. The $NMAD$ is reduced to 0.013, and the fractions of residuals for the stacked PDF, $f_{0.05}$ and, $f_{0.15}$, increase to, respectively, 85.9% and 99.3%. This was expected, since the role played by $NMAD$ and fraction of outliers for zphot point estimates, is analogous to the one of, respectively, $f_{0.05}$ and $f_{0.15}$ for the PDF. Another estimator that allows quantifying the reliability of the estimated PDF is the *zspecClass* flag as defined in Sect. 4. The results for *zspecClass* are reported in Table 4. As it could be expected, the best results in terms of fractions of *zspecClass* equal to 0 and 1, occur for the data sets *Cut-2* and *Cut-4*, which include the best scores in terms of zphot point estimates and of *cumulative* PDF performances (Table 3).

For the *Cut-5* data set, we achieve for classes 0 and 1 the scores of 31.0% and 42.8% respectively, and a smaller fraction of objects of class 3 (only 0.1%) with respect to the other data sets. Classes 3 and 4 quantify the number of objects falling outside the PDF. The distinction between the two classes gives the supplementary information about how far from the PDFs is their zspec.

In Fig. 7 we present the scatter plots of the zphot *best-estimates* as a function of the spectroscopic redshifts, for the test set and the four probed cut data sets. The mean and standard deviation of zphot are also plotted, in 40 evenly spaced zspec bins in a whole range of [0, 4.0]. Not all bins are populated, due to the reduction of the amount of samples resulting from the application of the rejection criteria, and the σ value in each bin increases in under-sampled bins.

Table 3 Statistics of the zphot and stacked PDFs for the whole test set sample and the four pruned data sets performed

Estimator	Test set	Cut 1	Cut 2	Cut 3	Cut 4
$bias$	−0.003	−0.004	−0.002	−0.003	−0.002
$NMAD$	0.016	0.015	0.015	0.015	0.015
$outliers$	2.4%	1.4%	0.6%	1.5%	0.6%
$f_{0.05}$	76.1%	79.2%	79.8%	77.5%	81.2%
$f_{0.15}$	94.6%	96.6%	97.4%	95.9%	98.0%
$\langle \Delta z \rangle$	−0.025	−0.016	−0.018	−0.008	−0.013

Table 4 *zspecClass* fractions for the whole test set and the four cuts.

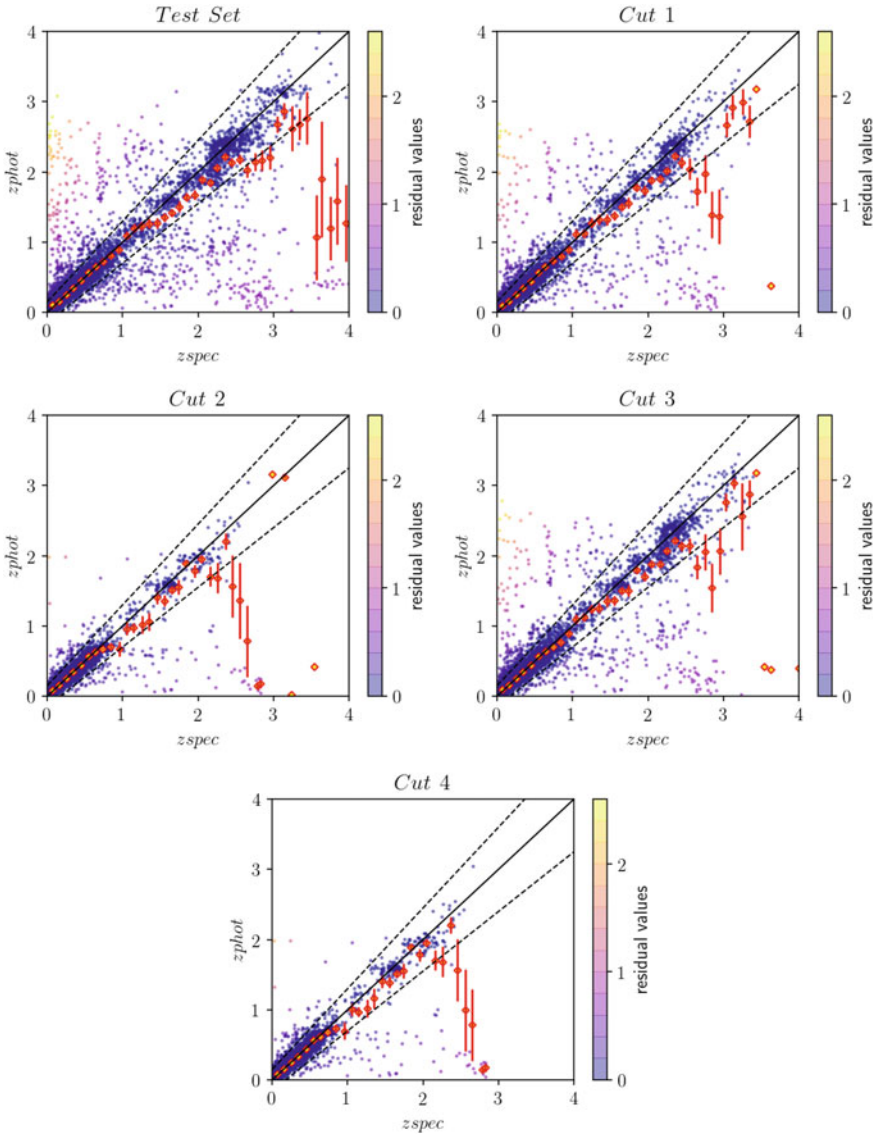| zspecClass	Test set		Cut 1		Cut 2		Cut 3		Cut 4	
0	10652	(26.1%)	9930	(27.5%)	10535	(27.6%)	10631	(26.6%)	9834	(28.3%)
1	15476	(37.9%)	14224	(39.3%)	15214	(39.9%)	15430	(38.7%)	14084	(40.5%)
2	13893	(34.0%)	11353	(31.4%)	11727	(30.8%)	13115	(32.9%)	10271	(29.5%)
3	156	(0.4%)	90	(0.2%)	79	(0.2%)	115	(0.3%)	64	(0.2%)
4	619	(1.5%)	600	(1.7%)	552	(1.4%)	614	(1.5%)	540	(1.6%)

Fig. 7 Scatter plots of photometric redshift *best-estimates* as a function of spectroscopic redshifts. Red diamonds: mean of the zphot *best-estimates* in 40 evenly spaced zspec bins. Red bars: standard deviation of zphot values populating the bins. Outliers and non-outliers are identified by the colour bar, showing the absolute values of the residuals (see Eq. 1)

It is interesting to notice the similar trends for data sets deriving from cuts 1 and 3, and cuts 2 and 4. This is expected for cut data sets 1 and 3 since, as we mentioned, PDF width and the number of bins in which PDF differ from 0, are highly correlated. In the case of cut data sets 2 and 4, being the cut data set 4 obtained by the joint application of cuts 1 and 2 conditions, the similar performance can shed light on the cut which drives the statistical performance outcome.

Of course, we should favour those rejection criteria that, leading to similar statistical performance, remove a smaller number of sources from the original data set. In other words, we should adopt rejection criteria corresponding to the best trade-off between precision and completeness.

Among the tested cuts, the best results in terms of precision and completeness are achieved for the data set *Cut-2*, which while showing comparable results (cf. Tables 3 and 4) to data set *Cut-4*, contains $\simeq 9\%$ more sources.

Finally, in Fig. 8, we show the stacked PDF for the test set and two cut data sets (*Cut-1* and *Cut-4*), along with the spectro-photometric redshift distributions of the test set. Note that the redshift range has been cut at $z = 1$, due to the low amount of objects in the test set above such value.

Besides the good agreement with the zphot distribution for the stacked PDF of the test set, we can see the effect of the rejection for the clipped data sets, which leads to a lower amount of sources at high redshift, and a more substantial amount at low redshift.
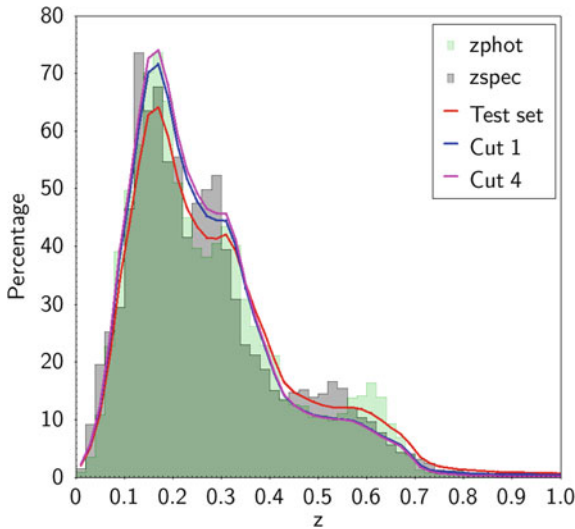

Fig. 8 Superposition of the stacked PDF (percentage) of the test set (red), for the *cut 1* (blue) and *cut 4* (magenta) data sets, and the zphot *best-estimates* (in green) distributions obtained by METAPHOR applied to the zspec distribution (in black), for the test set, limited to $z = 1$, due to the very few objects over such value

This is highlighted in Fig. 9, where the distribution of zspec for the whole test set and the two clipped data sets *Cut-1* and *Cut-4* are shown. Moreover, in Fig. 13 in the Appendix, it is reported the zspec distribution for the whole test set against the zspec distributions for the tested four clipped data sets. As we can see more clearly from this figure, rejection is successful in removing outliers at higher redshift. Comparing the distributions for the *Cut-2* and *Cut-4* data sets, we can notice that the range of zspec between 3 and 4 is more populated in the case of *Cut-2* data set with respect to *Cut-4* data set. This leads to the conclusion that, as anticipated above, *Cut-2* achieves a better trade-off between completeness and precision with respect to *Cut-4*. This is a clue of the effectiveness of the rejection in removing most of outliers at high redshift, i.e. in a region of the parameter space where the density of the training points is lower.
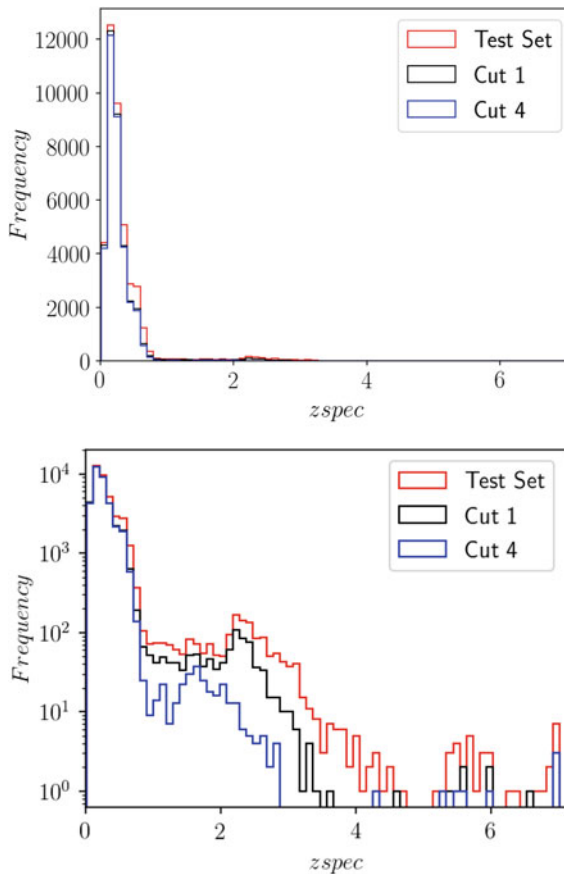

Fig. 9 *Top panel:* zspec distribution for the whole test set (red), and the cut data sets *Cut 1* (black) and *Cut 4* (blue). *Bottom panel*: the same of the top panel in a logarithmic scale
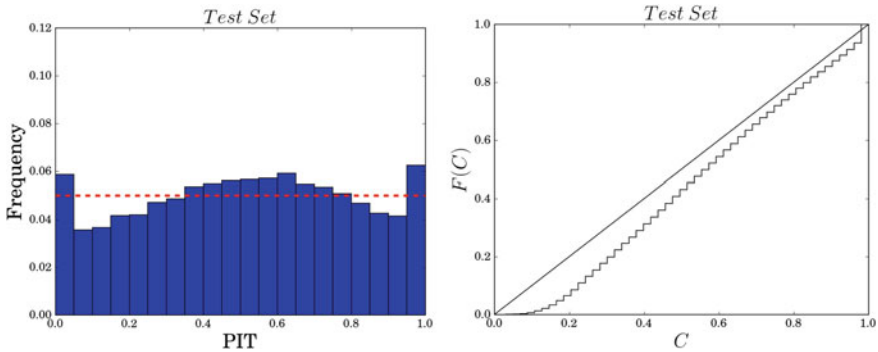
Fig. 10 PIT histogram (left panel). The red dashed line identifies the ideal PIT value, which represents the best calibration between the true zspec and the reconstructed zphot distributions. Credibility analysis (right panel) for the test set. The solid black line represents the best credibility for which the two distributions $F(C)$ and C are indistinguishable (see Sect. 4)

5.2 PIT and Credibility Analysis

PIT and credibility analysis for the test set are shown in Fig. 10. The PIT histogram shows a certain degree of *underdispersion* of the zphot distribution and the credibility plot stresses the *overconfidence* of the PDFs. The complementary information carried by these two *visual* diagnostics (see Sect. 4) is therefore confirmed.

In the plots of Figs. 11 and 12, we show, respectively, the credibility analysis for the four clipped data sets against the credibility of the test set, and the comparison of PIT and credibility for the data set *Cut-5*. Data sets obtained from cuts 1 and 4 show a slightly higher degree of *overconfidence* with respect to the test set, while cut data sets 2 and 3 show an indistinguishable credibility trend.

We stress that the PIT histogram fails to reveal differences between the four clipped data sets probed, with respect to the test set: for this reason, we do not show the relative plots. However, in the case of *Cut-5*, PIT shows a more significant degree of bias with respect to the test set, whereas the credibility shows a narrower shape, resulting in a larger *overconfidence* with respect to the whole test set.

6 Conclusions

In this work, we presented a method for defining low-quality zphot rejection criteria through the characterization of outliers, using the descriptors of the PDF shape (e.g. the width, the value of the maximum peak, etc.). The first step was, therefore, to compare the PDF descriptors for the two populations of outliers and non-outliers. Outliers appear to be characterized by wider PDFs with small maximum probability, as well as by a more significant number of bins in which the PDF differs from zero.
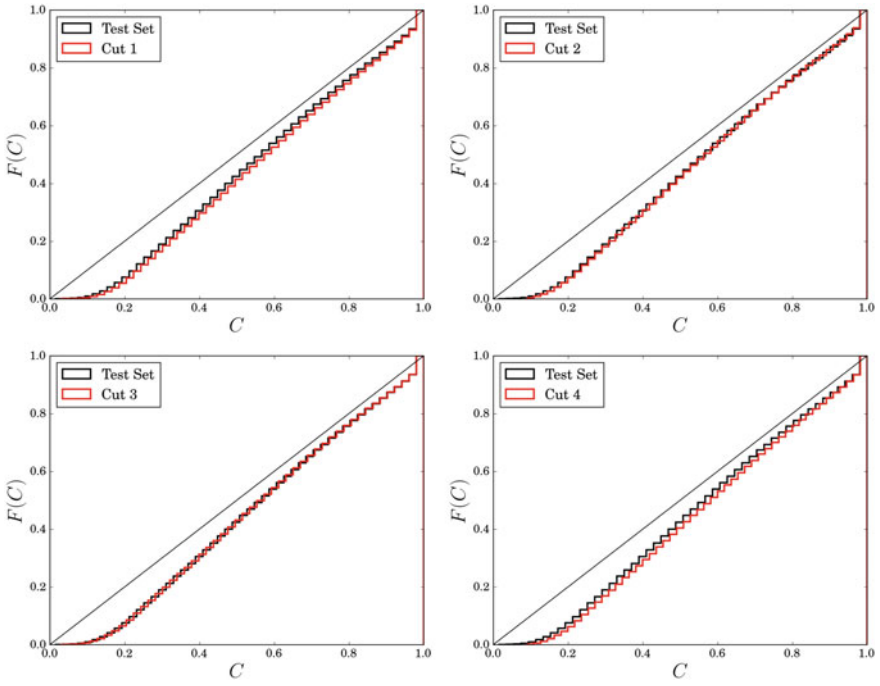
Fig. 11 Credibility analysis for the four tested cut data sets (red) against the whole test set credibility (black). In the top panels: *Cut-1* (left) and *Cut-2* (right). In the bottom panels: *Cut-3* (left) and *Cut-4* (right)
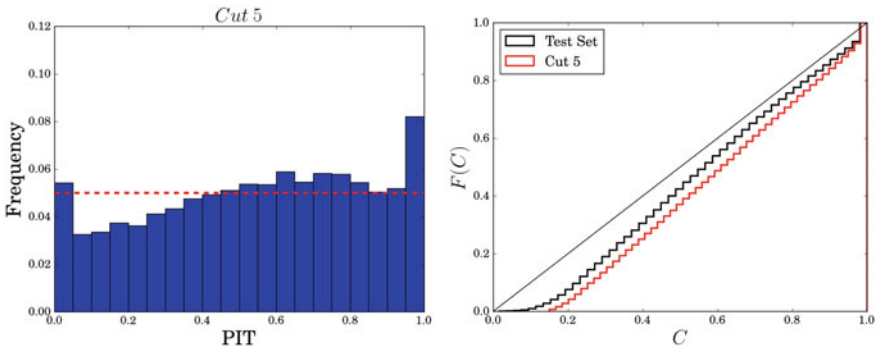

Fig. 12 PIT histogram (left panel) and credibility analysis (right panel) for the *Cut-5* data set, plotted against that of the whole test set

Zphot outliers tend to populate particular regions of the photometric parameter space and of the one defined by the PDF characteristics. Most outliers populate the top right part of a plot *PdfNBins* vs *PdfWidth*, where both the quantities are larger. Furthermore, the PDFs of the outliers have low maximum peaks, and populate a stripe

at low values of *PdfPeakHeight*, in a plane *PdfPeakHeight* vs *PdfWidth*. This allows the identification of cuts suitable to remove outliers, thus improving the precision on the clipped data sets.

We detailed the results for four different cut data sets obtained by applying rejection through the PDF width, the height of the maximum peak, and the number of bins in which the PDF is not null for, respectively *Cut-1, 2, 3* data sets. A further clipped data set (*Cut-4)* was created by removing outliers through the application of both the cuts used to generate *Cut-1* and *Cut-2* data sets. The best precision and completeness results were achieved for the data set *Cut-2*. This data set, in fact, from the one hand, shows comparable results to data set *Cut-4*, in terms of both zphot point estimate and *cumulative* PDF statistical performances. On the other hand, *Cut-2* data set contains $\simeq$ 9% more sources than *Cut-4*, which mostly populate the spectroscopic region in the range [3, 4], as it is visible in Fig. 13.

Finally, we tested many others more strict rejection criteria, all of them leading to a severe loss of completeness with respect to the original data set. We reported for one of these pruned data set (*Cut-5*) the results throughout the Sect. 5, also showing the more biased PIT histogram and more *overconfident* credibility diagram with respect to the other four pruned data sets (see Fig. 12).

Although still not fully automated, the rejection approach is very general in its applicability, since it does not depend on the particular method used to calculate the PDFs. On the other hand, the overall quality of PDFs depends strictly on the particular method used to derive them. This last aspect is not discussed in this paper. However, we deem particularly useful a future comparison of rejections applied to PDFs obtained by different approaches (e.g. SED and ML methods referenced in Sect. 2). This with the final goal of further increasing the precision of the measurements.

As mentioned in Sect. 1, precision and completeness are both relevant quantities for matching the requirements of ongoing as well as future cosmological sky surveys, since the accuracy of the cosmological parameters strongly depends on an optimal trade-off between these two properties. The systematic study and automatic implementation of rejection can help to improve the precision, keeping a congruous number of non-outliers objects, thus preserving the completeness.
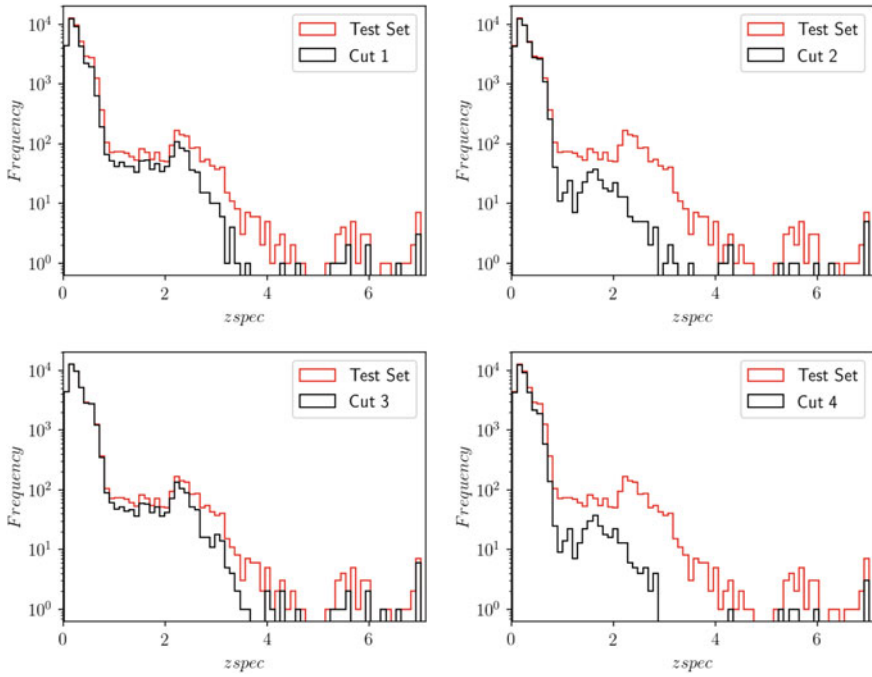
Fig. 13 Zspec distribution for the whole test set (red) and the four tested cut data sets (black), in a logarithmic scale. *Top panels*: *Cut-1* (left) and *Cut-2* (right). *Bottom panels*: *Cut-3* (left) and *Cut-4* (right)

Appendix

See Fig. 13

References

1. Serjeant, S.: ApJ **793**(1), L10 (2014). https://doi.org/10.1088/2041-8205/793/1/L10
2. Hildebrandt, H., Viola, M., Heymans, C., Joudaki, S., Kuijken, K., et al.: MNRAS **465**(2), 1454 (2017). https://doi.org/10.1093/mnras/stw2805
3. Fu, L., Liu, D., Radovich, M., Liu, X., Pan, C., et al.: Mon. Not. R. Astron. Soc. **479**(3), 3858 (2018). https://doi.org/10.1093/mnras/sty1579
4. Aragon-Calvo, M.A., Weygaert, R.v.d., Jones, B.J.T., Mobasher, B.: Mon. Not. R. Astron. Soc. **454**(1), 463 (2015). https://doi.org/10.1093/mnras/stv1903
5. Capozzi, D., de Filippis, E., Paolillo, M., D'Abrusco, R., Longo, G.: MNRAS **396**(2), 900 (2009). https://doi.org/10.1111/j.1365-2966.2009.14738.x
6. Annunziatella, M., Mercurio, A., Biviano, A., Girardi, M., Nonino, M., et al.: A&A **585**, A160 (2016). https://doi.org/10.1051/0004-6361/201527399

7. Radovich, M., Puddu, E., Bellagamba, F., Roncarelli, M., Moscardini, L., et al.: A&A **598**, A107 (2017). https://doi.org/10.1051/0004-6361/201629353

8. Brescia, M., Cavuoti, S., Paolillo, M., Longo, G., Puzia, T.: Mon. Not. R. Astron. Soc. **421**(2), 1155 (2012). https://doi.org/10.1111/j.1365-2966.2011.20375.x

9. Tortora, C., La Barbera, F., Napolitano, N.R., Roy, N., Radovich, M., et al.: Mon. Not. R. Astron. Soc. **457**(3), 2845 (2016). https://doi.org/10.1093/mnras/stw184

10. Masters, D., Capak, P., Stern, D., Ilbert, O., Salvato, M., et al.: ApJ **813**(1), 53 (2015). https://doi.org/10.1088/0004-637X/813/1/53

11. Gorecki, A., Abate, A., Ansari, R., Barrau, A., Baumont, S., Moniez, M., Ricol, J.S.: A&A **561**, A128 (2014). https://doi.org/10.1051/0004-6361/201321102

12. Ross, A.J., Banik, N., Avila, S., Percival, W.J., Dodelson, S., et al.: Mon. Not. R. Astron. Soc. **472**(4), 4456 (2017). https://doi.org/10.1093/mnras/stx2120

13. Baum, W.A.: In: Problems of Extra-Galactic Research, IAU Symposium, vol. 15, ed. by G.C. McVittie, p. 390 (1962)

14. Connolly, A.J., Csabai, I., Szalay, A.S., Koo, D.C., Kron, R.G., Munn, J.A.: AJ **110**, 2655 (1995). https://doi.org/10.1086/117720

15. Bolzonella, M., Miralles, J.M., Pelló, R.: A&A **363**, 476 (2000)

16. Arnouts, S., Cristiani, S., Moscardini, L., Matarrese, S., Lucchin, F., Fontana, A., Giallongo, E.: MNRAS **310**(2), 540 (1999). https://doi.org/10.1046/j.1365-8711.1999.02978.x

17. Tanaka, M.: Astrophys. J. **801**(1), 20 (2015). https://doi.org/10.1088/0004-637x/801/1/20

18. Tagliaferri, R., Longo, G., Andreon, S., Capozziello, S., Donalek, C., Giordano, G.: arXiv:astro-ph/0203445, https://doi.org/10.1007/978-3-540-45216-4_26 (2000)

19. Firth, A.E., Lahav, O., Somerville, R.S.: Mon. Not. R. Astron. Soc. **339**(4), 1195 (2003). https://doi.org/10.1046/j.1365-8711.2003.06271.x

20. Ball, N.M., Brunner, R.J., Myers, A.D., Strand, N.E., Alberts, S.L., Tcheng, D.: ApJ **683**(1), 12 (2008). https://doi.org/10.1086/589646

21. Carrasco Kind, M., Brunner, R.J.: Astronomical data analysis software and systems XXII. In: D.N. Friedel (ed.) Astronomical Society of the Pacific Conference Series, vol. 475, p. 69. Astronomical Society of the Pacific, San Francisco (2013)

22. Brescia, M., Cavuoti, S., Longo, G., De Stefano, V.: A&A **568**, A126 (2014). https://doi.org/10.1051/0004-6361/201424383

23. Graff, P., Feroz, F., Hobson, M.P., Lasenby, A.: MNRAS **441**(2), 1741 (2014). https://doi.org/10.1093/mnras/stu642

24. Cavuoti, S., Brescia, M., Tortora, C., Longo, G., Napolitano, N.R., et al.: MNRAS **452**(3), 3100 (2015). https://doi.org/10.1093/mnras/stv1496

25. Cavuoti, S., Brescia, M., De Stefano, V., Longo, G.: Experim. Astron. **39**(1), 45 (2015). https://doi.org/10.1007/s10686-015-9443-4

26. Sadeh, I., Abdalla, F.B., Lahav, O.: PASP **128**(968), 104502 (2016). https://doi.org/10.1088/1538-3873/128/968/104502

27. Soo, J.Y.H., Moraes, B., Joachimi, B., Hartley, W., Lahav, O., et al.: MNRAS **475**(3), 3613 (2018). https://doi.org/10.1093/mnras/stx3201

28. D'Isanto, A., Cavuoti, S., Gieseke, F., Polsterer, K.L.: A&A **616**, A97 (2018). https://doi.org/10.1051/0004-6361/201833103

29. Fluke, C.J., Jacobs, C.: WIREs Data Min. Knowl. Discov. **10**(2), e1349 (2020). https://doi.org/10.1002/widm.1349

30. Hildebrandt, H., Arnouts, S., Capak, P., Moustakas, L.A., Wolf, C., et al.: A&A **523**, A31 (2010). https://doi.org/10.1051/0004-6361/201014885

31. Abdalla, F.B., Banerji, M., Lahav, O., Rashkov, V.: Mon. Not. R. Astron. Soc. **417**(3), 1891 (2011). https://doi.org/10.1111/j.1365-2966.2011.19375.x

32. Sánchez, C., Carrasco Kind, M., Lin, H., Miquel, R., Abdalla, F.B., et al., MNRAS **445**(2), 1482 (2014). https://doi.org/10.1093/mnras/stu1836

33. Cavuoti, S., Tortora, C., Brescia, M., Longo, G., Radovich, M., et al.: MNRAS **466**(2), 2039 (2017). https://doi.org/10.1093/mnras/stw3208

34. Schmidt, S.J., Malz, A.I., Soo, J.Y.H., Almosallam, I.A., Brescia, M., et al.: arXiv:2001.03621 (2020)
35. Brescia, M., Cavuoti, S., Amaro, V., Riccio, G., Angora, G., Vellucci, C., Longo, G.: arXiv:1802.07683 (2018)
36. Amaro, V., Cavuoti, S., Brescia, M., Vellucci, C., Longo, G., et al.: MNRAS **482**(3), 3116 (2019). https://doi.org/10.1093/mnras/sty2922
37. Mandelbaum, R., Seljak, U., Hirata, C.M., Bardelli, S., Bolzonella, M., et al.: MNRAS **386**(2), 781 (2008). https://doi.org/10.1111/j.1365-2966.2008.12947.x
38. Viola, M., Cacciato, M., Brouwer, M., Kuijken, K., Hoekstra, H., et al.: MNRAS **452**(4), 3529 (2015). https://doi.org/10.1093/mnras/stv1447
39. Brammer, G.B., van Dokkum, P.G., Coppi, P.: ApJ **686**(2), 1503 (2008). https://doi.org/10.1086/591786
40. Ilbert, O., Arnouts, S., McCracken, H.J., Bolzonella, M., Bertin, E., et al.: A&A **457**(3), 841 (2006). https://doi.org/10.1051/0004-6361:20065138
41. Benítez, N.: ApJ **536**(2), 571 (2000). https://doi.org/10.1086/308947
42. Bonnett, C.: MNRAS **449**(1), 1043 (2015). https://doi.org/10.1093/mnras/stv230
43. Carrasco Kind, M., Brunner, R.J.: MNRAS **432**(2), 1483 (2013). https://doi.org/10.1093/mnras/stt574
44. Carrasco Kind, M., Brunner, R.J.: MNRAS **438**(4), 3409 (2014). https://doi.org/10.1093/mnras/stt2456
45. Carrasco Kind, M., Brunner, R.J.: MNRAS **442**(4), 3380 (2014). https://doi.org/10.1093/mnras/stu1098
46. Kuijken, K., Heymans, C., Dvornik, A., Hildebrandt, H., de Jong, J.T.A., et al.: A&A **625**, A2 (2019). https://doi.org/10.1051/0004-6361/201834918
47. Edge, A., Sutherland, W., Kuijken, K., Driver, S., McMahon, R., Eales, S., Emerson, J.P.: Messenger **154**, 32 (2013)
48. Davies, L.J.M., Robotham, A.S.G., Driver, S.P., Alpaslan, M., Baldry, I.K., et al.: MNRAS **452**(1), 616 (2015). https://doi.org/10.1093/mnras/stv1241
49. Lilly, S.J., Brun, V.L., Maier, C., Mainieri, V., Mignoli, M., et al.: Astrophys. J. Suppl. Ser. **184**(2), 218 (2009). https://doi.org/10.1088/0067-0049/184/2/218
50. Cooper, M.C., Yan, R., Dickinson, M., Juneau, S., Lotz, J.M., et al.: Mon. Not. R. Astron. Soc. **425**(3), 2116 (2012). https://doi.org/10.1111/j.1365-2966.2012.21524.x
51. Newman, J.A., Cooper, M.C., Davis, M., Faber, S.M., Coil, A.L., et al.: ApJS **208**(1), 5 (2013). https://doi.org/10.1088/0067-0049/208/1/5
52. Liske, J., Baldry, I.K., Driver, S.P., Tuffs, R.J., Alpaslan, M., et al.: MNRAS **452**(2), 2087 (2015). https://doi.org/10.1093/mnras/stv1436
53. Baldry, I.K., Liske, J., Brown, M.J.I., Robotham, A.S.G., Driver, S.P., et al.: MNRAS **474**(3), 3875 (2018). https://doi.org/10.1093/mnras/stx3042
54. Cavuoti, S., Amaro, V., Brescia, M., Vellucci, C., Tortora, C., Longo, G.: MNRAS **465**(2), 1959 (2017). https://doi.org/10.1093/mnras/stw2930
55. Brescia, M., Cavuoti, S., D'Abrusco, R., Longo, G., Mercurio, A.: ApJ **772**(2), 140 (2013). https://doi.org/10.1088/0004-637X/772/2/140
56. Brescia, M., Cavuoti, S., Longo, G., Nocella, A., Garofalo, M., et al.: PASP **126**(942), 783 (2014). https://doi.org/10.1086/677725
57. Wittman, D., Bhaskar, R., Tobin, R.: MNRAS **457**(4), 4005 (2016). https://doi.org/10.1093/mnras/stw261
58. Gneiting, T., Balabdaoui, F., Raftery, A.E.: J. R. Stat. Soc. Ser. B **69**(2), 243 (2007). https://EconPapers.repec.org/RePEc:bla:jorssb:v:69:y:2007:i:2:p:243-268